Cooccurrence analysis between high and low-rank tags

Fig. 7 shows a table where the occurrence of 30 high-rank (low-frequency) tags is related to the occurrence of the 15 lowest-rank (highest-frequency) tags. All the tags under study are cooccurring with the tag *blog* and the dataset used for the analysis is the same as the one used in Fig. 2. The cooccurrence analysis is performed as follows: given a high-rank tag X, all resources tagged with X (within the above dataset) are selected, and the cooccurrence frequencies of X with each of the 15 top-ranked (most frequent) tags are recorded. Thus, each row of the table associates a tag X with the corresponding (normalized) cooccurrence histogram. This provides a statistical characterization of tag X in terms of the top-ranked tags, regarded as a natural basis for categorization (or semantic "grounding"). Fig. 8 graphically illustrates such a "tag fingerprint" for 5 high-rank tags, arbitrarily chosen. This analysis is aimed at probing the existence of nontrivial cooccurrence relationships that might be ascribed to semantics and – possibly – to the emergence of a self-organized hierarchy of tags. As shown by the bold numbers in Fig. 7, as well as by the graph in Fig. 8, high-frequency (low-rank) tags do not trivially cooccur with most of the low-frequency (high-rank) tags — on the contrary, the cooccurrence profile of the latter is peaked in correspondence of specific, semantically related tags (*economics* and *law* with *politics*, for example, see Fig. 8). Moreover, several low-frequency (high-rank) tags never cooccur with some of the highest-frequency (low-rank) tags, as shown by the several zeros in Fig. 7. This suggests that high-frequency tags partition – or "categorize" – the resources marked by tags of lower frequency. Given that our definitions of "high-rank" and "low-rank" are somehow arbitrary, and given the self-similar character of tag association we observed (Fig. 3), we expect our observations to be representative of a general and complex semiotic structure underlying folksonomies.

	design	web	news	music	rss	css	daily	art	politics	tech	technology	blogs	software	media	programming
inspiration	**0.392**	0.159	0.026	0.011	**0.000**	0.206	0.011	0.132	**0.000**	0.011	0.021	0.032	**0.000**	**0.000**	**0.000**
socialsoftware	0.066	0.242	0.022	0.011	0.099	0.044	**0.000**	0.022	**0.000**	0.055	0.055	0.121	0.176	0.077	0.011
economics	**0.000**	0.119	0.090	**0.000**	**0.000**	**0.000**	0.119	**0.000**	**0.552**	0.030	0.045	0.030	**0.000**	0.015	**0.000**
opensource	0.081	0.144	0.108	0.018	0.063	0.009	0.027	0.009	**0.000**	0.054	0.045	0.090	0.243	0.036	0.072
computer	0.167	0.080	0.093	0.013	0.027	0.073	0.033	0.020	0.020	0.080	0.133	0.060	0.127	0.007	0.067
python	0.031	0.138	0.077	**0.000**	0.092	0.046	**0.000**	**0.000**	**0.000**	0.015	0.046	**0.000**	0.092	**0.000**	**0.462**
tagging	0.234	0.213	0.021	**0.000**	0.085	0.085	0.021	**0.000**	0.021	0.021	0.085	0.043	0.170	**0.000**	**0.000**
comics	0.089	0.089	0.089	0.054	**0.000**	0.018	0.054	**0.482**	0.018	**0.000**	0.036	0.054	**0.000**	0.018	**0.000**
research	0.096	0.135	0.096	**0.000**	0.058	**0.000**	0.058	0.038	0.038	0.019	0.154	0.077	0.096	0.096	0.038
law	0.012	0.061	0.134	0.037	**0.000**	**0.000**	0.110	**0.000**	**0.378**	0.024	0.195	0.049	**0.000**	**0.000**	**0.000**
hack	0.080	0.080	0.069	0.034	0.057	0.046	0.126	0.023	0.034	0.161	0.103	0.057	0.069	0.023	0.034
xhtml	**0.311**	**0.262**	0.016	0.004	0.016	**0.336**	**0.000**	0.004	**0.000**	0.008	0.012	0.004	0.012	0.004	0.008
humour	0.125	0.042	0.125	**0.000**	**0.000**	**0.000**	0.125	0.208	0.083	0.083	0.042	**0.000**	**0.000**	0.042	0.125
management	0.147	0.191	**0.000**	**0.000**	0.029	0.015	0.118	0.015	**0.000**	0.015	0.103	0.103	0.162	0.029	0.074
movies	0.080	0.080	0.240	0.120	**0.000**	**0.000**	0.100	0.140	0.080	**0.000**	0.020	0.100	**0.000**	0.040	**0.000**
diy	0.151	0.081	0.047	0.023	**0.000**	**0.000**	0.081	0.186	0.012	0.198	0.093	0.047	0.058	0.012	0.012
life	0.231	**0.000**	**0.000**	0.019	**0.000**	**0.000**	0.096	0.250	0.058	0.058	0.250	**0.000**	0.019	**0.000**	0.019
tag	0.125	**0.354**	**0.000**	**0.000**	0.250	0.021	**0.000**	0.021	**0.000**	0.021	0.062	0.021	0.083	0.021	0.021
maps	0.200	0.150	0.050	**0.000**	0.050	**0.000**	**0.000**	0.050	0.050	0.250	0.050	0.050	0.050	0.050	**0.000**
ideas	0.167	0.103	0.090	**0.000**	0.026	0.077	0.077	0.051	0.038	0.064	0.103	0.077	0.051	0.064	0.013
architecture	**0.526**	0.063	0.074	0.011	**0.000**	**0.000**	0.095	0.137	**0.000**	**0.000**	0.032	0.011	0.011	**0.000**	0.042
organization	**0.000**	0.042	**0.000**	0.021	**0.000**	**0.000**	**0.479**	0.021	**0.000**	0.062	0.125	0.021	0.146	0.042	0.042
plugin	0.156	**0.281**	**0.000**	0.031	0.031	0.031	**0.000**	**0.000**	**0.000**	**0.000**	0.031	0.031	**0.312**	0.031	0.062
blogroll	0.054	0.089	0.107	0.071	0.054	**0.000**	0.054	0.054	0.071	0.089	0.107	0.125	0.054	0.018	0.054
information	0.065	**0.280**	0.172	**0.000**	0.097	0.022	0.022	0.011	0.043	0.054	0.075	0.054	0.054	0.032	0.022
articles	0.159	0.136	0.091	0.011	**0.000**	0.068	0.057	0.011	0.091	0.057	0.080	0.045	0.091	0.045	0.057
resource	**0.287**	0.202	0.032	0.032	0.032	0.191	0.021	0.032	0.011	0.053	0.011	0.011	0.032	**0.000**	0.053
illustration	**0.339**	0.104	0.009	0.026	**0.000**	0.078	0.017	**0.400**	**0.000**	**0.000**	**0.000**	0.017	0.009	**0.000**	**0.000**
mobile	0.021	0.085	0.106	0.021	0.021	**0.000**	0.128	0.106	**0.000**	0.106	**0.255**	0.043	0.064	0.043	**0.000**
liberal	**0.000**	**0.000**	0.125	**0.000**	**0.000**	**0.000**	0.067	**0.000**	**0.692**	**0.000**	**0.000**	0.058	**0.000**	0.058	**0.000**

Figure 7: Cooccurrence table: columns correspond to the 15 top-ranked tags cooccurring with *blog*, in descending order of frequency from left to right. Rows correspond to 30 low-frequency tags cooccurrinng with *blog* (frequencies ranking between 100th and 200th). Each row is a normalized cooccurrence histogram representing a "categorization" of the corresponding tag in terms of the top-ranked tags. Numbers in red (bold face) denote cooccurrence probabilities in excess of 25%. Zeros (no cooccurrence) are marked in blue (bold face).

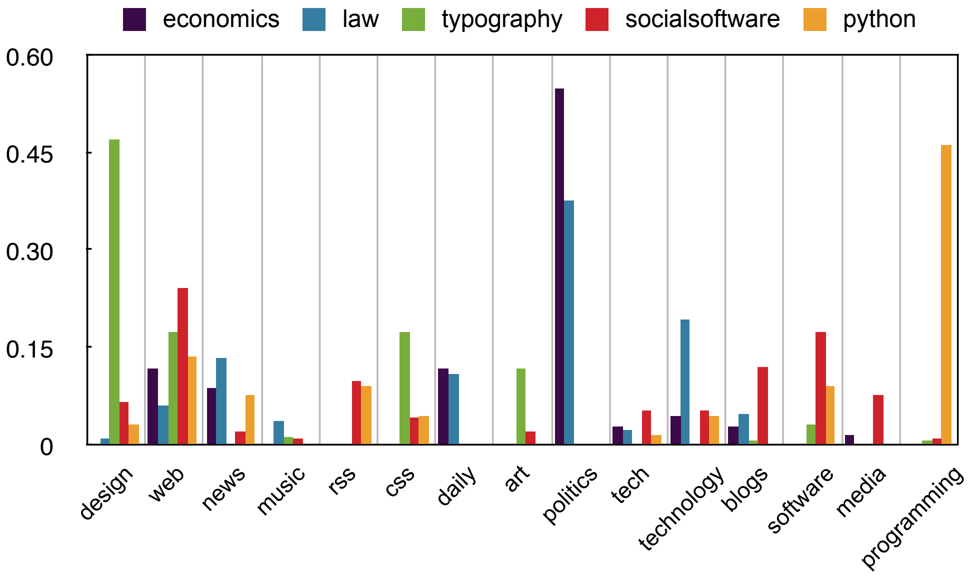

Figure 8: Cooccurrence patterns for 5 of the low-frequency (high-rank) tags of Fig. 7 (see legend at the top). The colored bars display the "fingerprint" of the selected tags in terms of their cooccurrence with the 15 top-ranked tags (the same ones reported in the top row of Fig. 7).

Role of the Parameters in the Yule-Simon Process with Memory

Here we investigate numerically the effect of the model parameters on the statistical properties of the simulated stream of tags, namely the the frequency-rank distribution $P(R)$ and the frequency probability distribution $P(k)$.

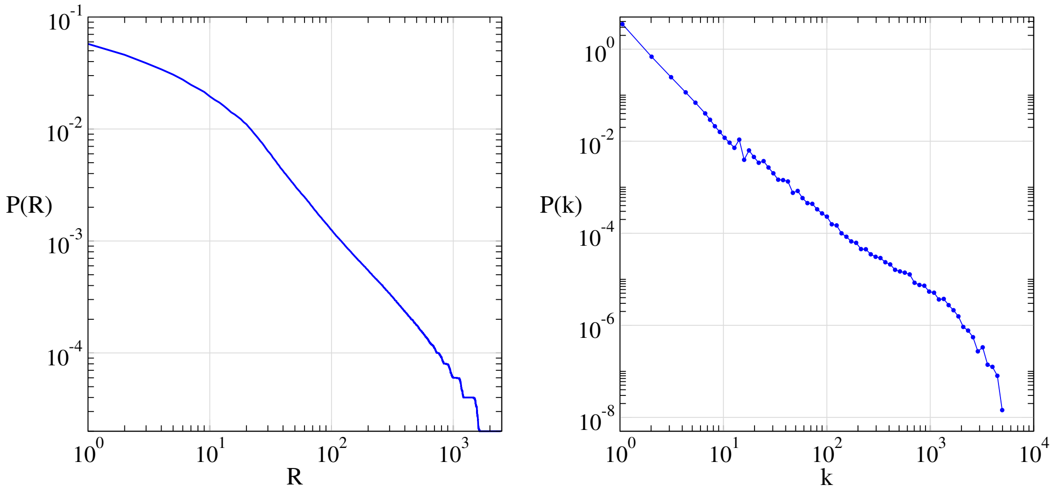

Figure 9: The behavior of our Yule-Simon process with a fat-tailed memory kernel is shown above for typical values of the model parameters: probability $p = 0.05$, memory parameter $\tau = 200$, initial number of words $n_0 = 50$, and a simulated time $t = 5 \cdot 10^4$. The frequency-rank distribution $P(R)$ (left panel) displays a power-law tail for high ranks and a low-rank flattening. The exponent of the power law is higher than 1, in contrast to the original Yule-Simon process, where its value is $1 - p$. The corresponding frequency probability distribution $P(k)$ (right panel) displays an overall power-law dependence with a sharp fall at high frequencies, corresponding to the low-rank flattening of the $P(R)$. All curves were computed by averaging 50 realizations of the process with the same set of parameter values.

For completeness we report also the experimental frequency distribution corresponding to the frequency-rank curves shown in the Fig. 2*b*. Here the noisy character of the last bins is even more evident than in the model.

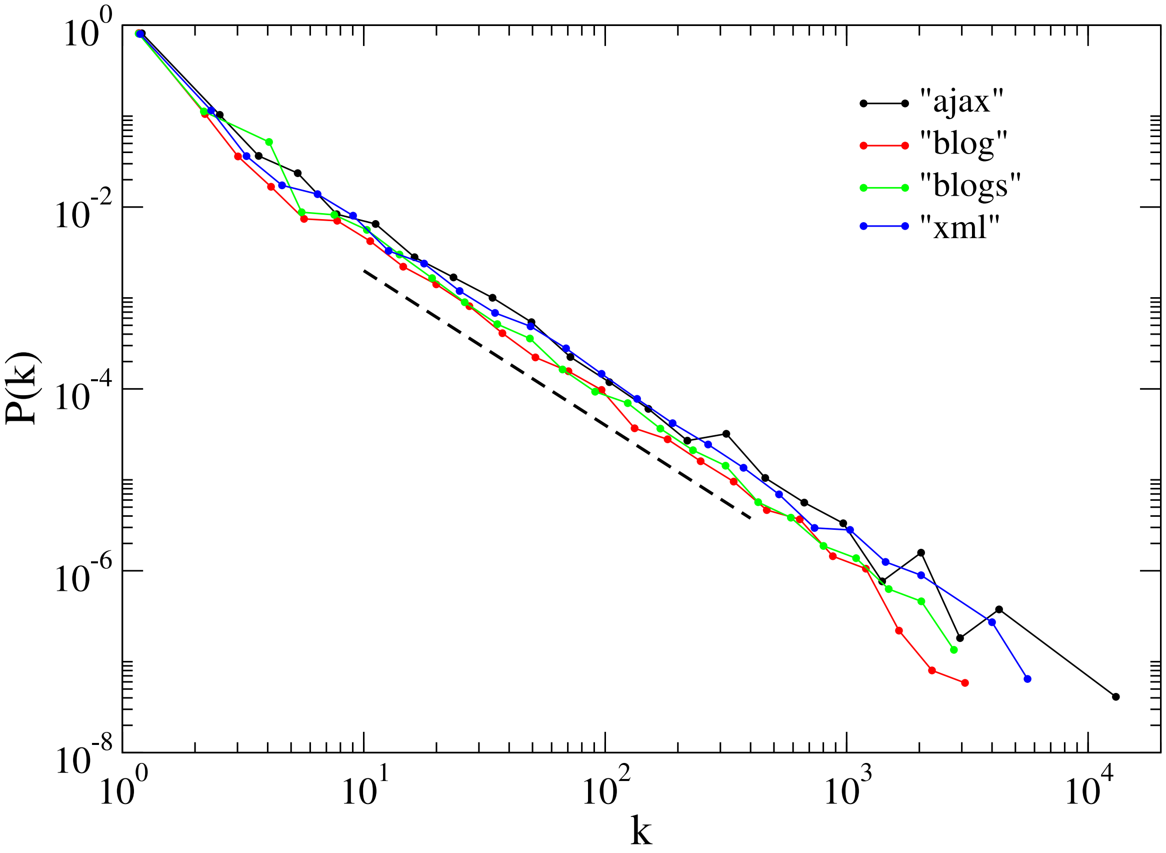

Figure 10: Tags frequency distribution corresponding to the frequency-rank curves shown in the Fig. 2*b* of the main text.

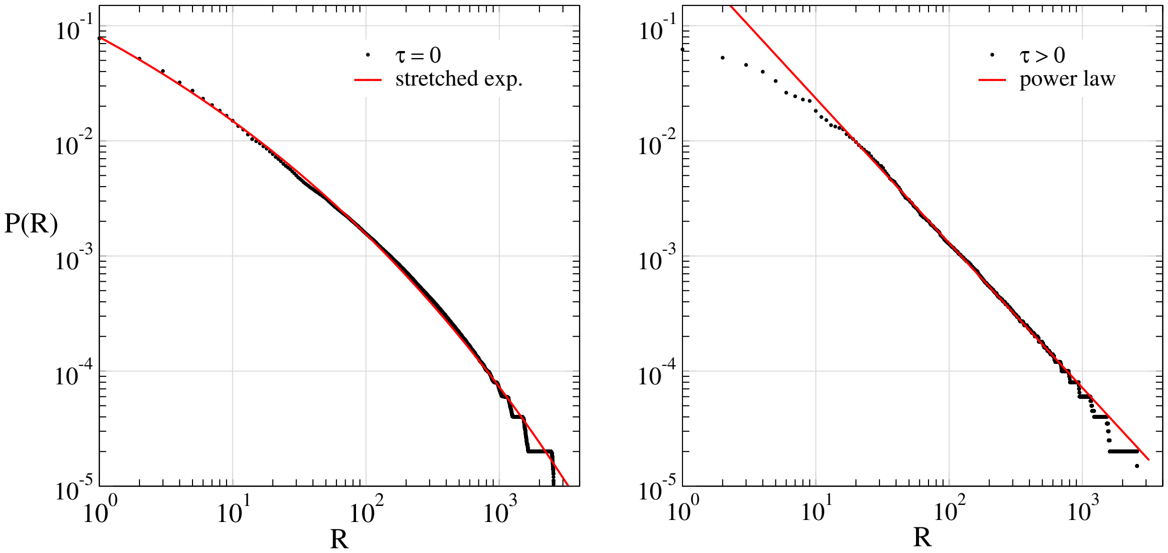

Figure 11: **Effect of τ** on the (normalized) frequency-rank distribution $P(R)$. For $\tau = 0$ (left panel), $P(R)$ displays a stretched exponential dependence on the rank R (red curve), as predicted by a mean-field master equations approach. For $\tau > 0$ (right panel, $\tau = 200$) the distribution develops a power-law tail (red line) and retains the low-rank flattening. τ also appears to control a crossover from the stretched exponential behavior we observe for $\tau = 0$ to the power-law tail behavior we observe for large τ The values of the parameters are $p = 0.05$, $n_0 = 50$, $t = 5 \cdot 10^4$. All curves were computed by averaging over 50 realizations of the process.

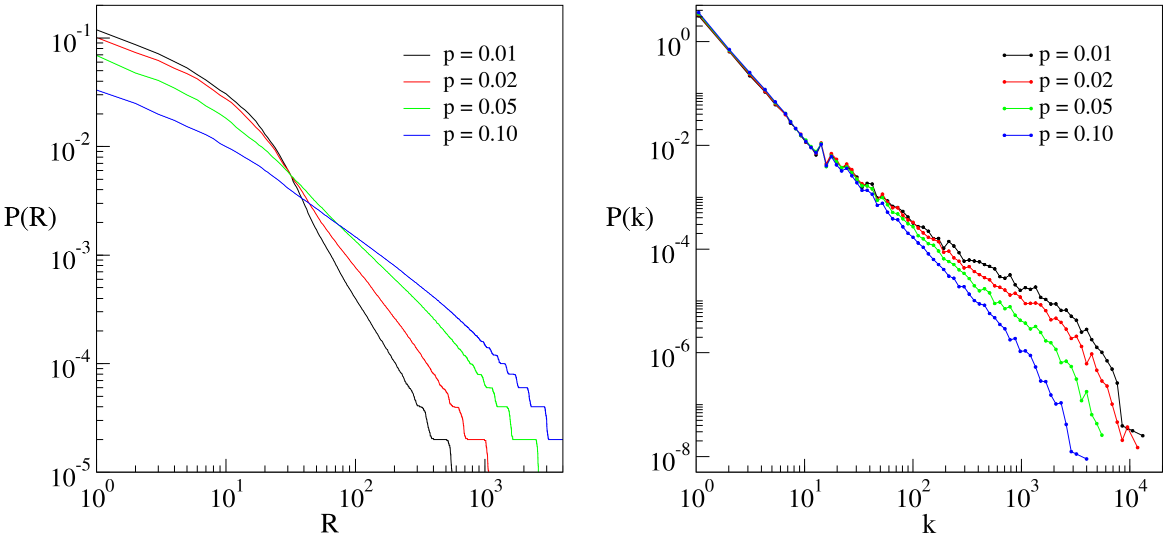

Figure 12: **Effect of *p*** on the (normalized) frequency-rank distribution $P(R)$ and on the frequency probability distribution $P(k)$, for a fixed value of $\tau = 50$. Just as in the simple Yule-Simon process, p affects the slope of the power-law tail of $P(R)$ (left panel). Conversely, the slope at low ranks (i.e. the low-rank flattening) is not significantly affected by the value of p. In terms of $P(k)$ (right panel), changes in p affect the high-frequency behavior of the distribution, with high-frequency tags becoming less and less likely for increasing values of p. Parameter values are $n_0 = 50$ and $t = 51000$, and a transient consisting of the first 1000 words was discarded. All curves were computed by averaging over 50 realizations of the process.

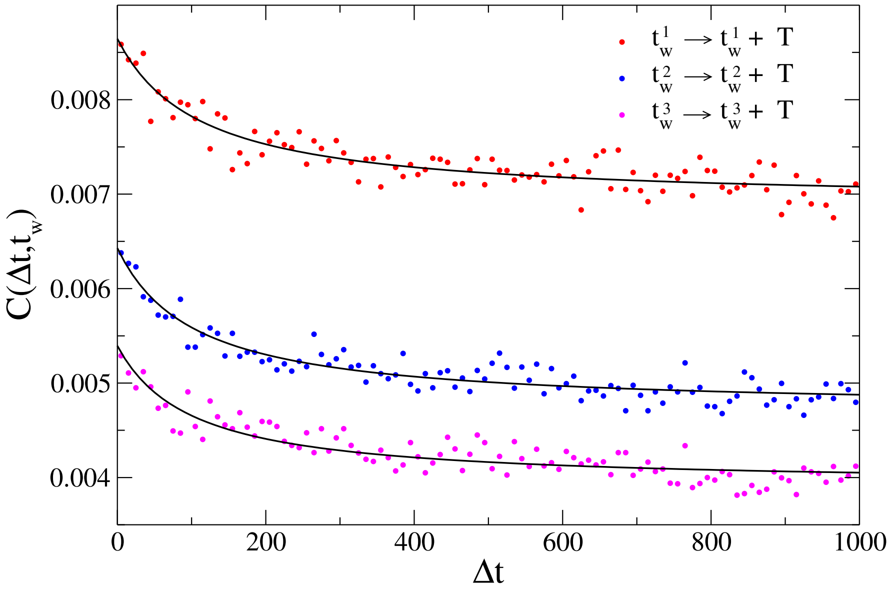

Figure 13: **Tag-tag correlations for simulated data.** The autocorrelation function $C(\Delta t, t_w)$ (see main text) is computed over three consecutive and equally long (30000 tags each) subsets of a simulated tag stream, starting respectively at $t_w = 1 \cdot 10^4$, $t_w = 4 \cdot 10^4$ and $t_w = 7 \cdot 10^4$. The values of the model parameters are $p = 0.06$, $\tau = 100$, $n_0 = 100$, $t = 1.1 \cdot 10^5$, and an initial transient of 10^4 tags was discarded. Short-range correlations are clearly visible, decaying towards a long-range plateau value. The solid black lines are obtained by fitting the simulated autocorrelation function with $C(\Delta t, t_w) = A_1 + A_2/(\Delta t + \tau)$, and show that the chosen form of the memory kernel induces the $1/\Delta t$ correlations we observe in experimental data (see main text). We want to remark that while the correlations in the simulated stream have the correct dependence on the lag Δt, the plateau value reached for $\Delta t \gg 1$ decreases with time, in contrast to what we experimentally observe. This difference can be understood by observing that for $\Delta t \gg 1$ we have $C(\Delta t, t_w) \simeq \sum_{R=1}^{R=R_{\max}(T,t_w)} P_{T,t_w}^2(R)$, so that the plateau value is sensitive to the total number of distinct tags, i.e. the maximum rank $R_{\max}$. Our simple model assumes that the rate of creation of new tags (p) is a constant, so that $R_{\max}$ increases linearly with the time t. On the other hand, we have experimental evidence (Fig. 14) that the growth of $R_{\max}$ is actually sub-linear, so that our model cannot possibly reproduce the correct time dependence of the plateau value of $C(\Delta t, t_w)$ (i.e. the long-term correlations).

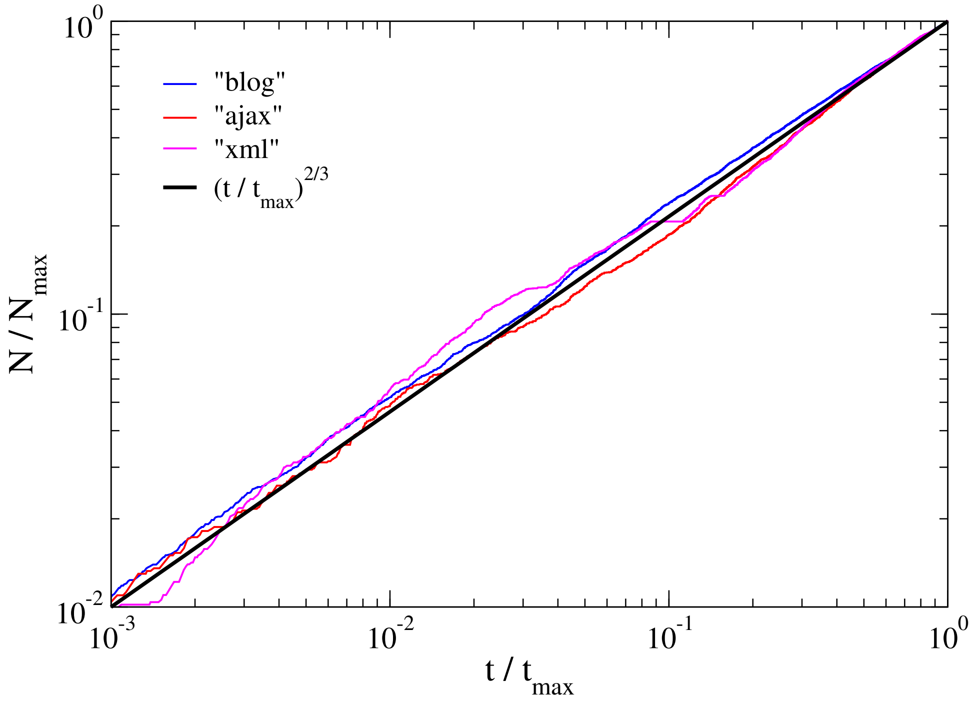

Figure 14: **Accumulation of tags.** We select a tag X and study how the number N of distinct tags co-occurring with X increases as a function of the total number t of tags co-occurring with X (the intrinsic "time" of the system). A sub-linear growth behavior can be observed for all the tags under study. Here we display the experimental data for three different tags (colored curves): on rescaling N and t by the values they assume at the end of the experimental time series, $N_{\max}$ and $t_{\max}$, all the curves collapse on the power-law $N/N\mathrm{max} = (t/t\mathrm{max})^{2/3}$ (solid black line). N never approaches a stationary value, while its rate of change decreases monotonically towards zero according to $dN/dt \sim t^{-1/3}$. The motivations of this universal behaviour are still unknown and they cannot be predicted in a self-consistent way in the framework of a purely statistical model (as ours) where only the stream statistics matters. That is why we preferred to keep our model as simple as possible not inserting any ad hoc assumption about the time-behaviour of N. For a deeper comprehension of the observed phenomenology a model embodying cognitive aspects is needed which is outside the scope of the present investigation.

Continuum Description of the Yule-Simon Process with Memory

The Model

We start with n_0 words. At a generic (discrete) time step t, a new word may be invented with probability p and appended to the text, while with probability $1 - p$ one word is copied from the text, going back in time by i steps with a probability that decays with i as a power law (see Fig. 3 in the main text),

$$Q(i) = \frac{C}{\tau + i} \, . \tag{1}$$

C is a time-dependent normalization factor and τ is a characteristic time-scale over which recently added words have comparable probabilities. The normalization condition for the memory kernel $Q(i)$ reads:

$$1 = \sum_{i=1}^{i=t} Q(i) = \sum_{i=1}^{i=t} \frac{C}{\tau + i} = C \sum_{i=1}^{i=t} \frac{1}{\tau + i} \ ,$$

so that

$$C(t) = \left(\sum_{i=1}^{i=t} \frac{1}{\tau + i} \right)^{-1} \, . \tag{2}$$

In the following we will write the normalization factor as C, with no explicit mention of its time dependence. We also define

$$\alpha(t) \equiv (1 - p) \, C(t) \, , \tag{3}$$

and we will similarly refer to it as α.

Return Times

We assume that word X occurred at time t for the first time, and we ask what is the probability $P(\Delta t)$ that the next occurrence of X happens at time $t + \Delta t$, with $\Delta t \geq 1$.

If $\Delta t = 1$, $P(\Delta t)$ is the probability of replicating the previous word, i.e. the product between the probability $1-p$ of copying an old word, and the probability of choosing the immediately preceding word ($i = 1$) computed according to the chosen memory kernel, $Q(1) = C/(\tau + 1)$. This gives

$$P(1) = (1 - p) \frac{C}{\tau + 1} = \frac{\alpha}{\tau + 1} \, . \tag{4}$$

For $\Delta t > 1$, $P(\Delta t)$ can be computed as the product of the probabilities of *not* choosing word X for $\Delta t - 1$ consecutive steps, multiplied by the probability of choosing word X at step Δt. In order not to choose word X at the first step, one has to either append a new word (probability p) or copy an existing word (probability $1-p$) which is not X (probability $1 - C/(\tau+1)$). Thus, the probability of not choosing word X at the first step is

$$p + (1-p)\left(1 - \frac{C}{\tau+1}\right) ,$$

and similarly the probability of not choosing word X at step i is

$$p + (1-p)\left(1 - \frac{C}{\tau+i}\right) ,$$

under the approximation that C is constant from step to step, i.e. $\Delta t \ll t$. Finally, under the same approximation, the probability of choosing word X at step $i = \Delta t$ is $(1-p)C/(\tau+\Delta t)$.

Putting everything together, for $\Delta t > 1$, we can write the return probability as the product

$$\begin{aligned} P(\Delta t) \;\; \simeq \;\; & \left[p + (1-p)\left(1 - \frac{C}{\tau+1}\right)\right] \cdot \\ & \left[p + (1-p)\left(1 - \frac{C}{\tau+2}\right)\right] \cdot \\ & \left[p + (1-p)\left(1 - \frac{C}{\tau+3}\right)\right] \cdot \\ & \cdots \\ & \cdot \left[p + (1-p)\left(1 - \frac{C}{\tau+\Delta t-1}\right)\right] \cdot \left[(1-p)\frac{C}{\tau+\Delta t}\right] . \end{aligned} \tag{5}$$

Taking the logarithm of $P(\Delta t)$, we can write the above product as the sum over steps $i = 1, 2, \ldots, \Delta t - 1$:

$$\begin{aligned} \ln P(\Delta t) \;\; &= \;\; \sum_{i=1}^{\Delta t-1} \ln\left[p + (1-p)\left(1 - \frac{C}{\tau+i}\right)\right] + \ln\frac{(1-p)\,C}{\tau+\Delta t} = \\ &= \;\; \sum_{i=1}^{\Delta t-1} \ln\left[1 - \frac{(1-p)\,C}{\tau+i}\right] + \ln\frac{(1-p)\,C}{\tau+\Delta t} = \\ &= \;\; \sum_{i=1}^{\Delta t-1} \ln\left(1 - \frac{\alpha}{\tau+i}\right) + \ln\frac{\alpha}{\tau+\Delta t} = \\ &\simeq \;\; -\alpha \sum_{i=1}^{\Delta t-1} \frac{1}{\tau+i} + \ln\frac{\alpha}{\tau+\Delta t} , \end{aligned} \tag{6}$$

where we used Eq. **3** and the fact that $\alpha \ll 1$ for $t \gg 1$.

Case $\tau = 0$

For $\tau = 0$, Eq. **6** becomes

$$\ln P(\Delta t) \simeq -\alpha \sum_{i=1}^{\Delta t-1} \frac{1}{i} + \ln \frac{\alpha}{\Delta t}\,, \tag{7}$$

which allows us to rewrite the harmonic sum in terms of the digamma function ψ_0:

$$\sum_{i=1}^{\Delta t-1} \frac{1}{i} = \gamma + \psi_0(\Delta t)\,, \tag{8}$$

where $\gamma = 0.577216\ldots$ is the Euler-Mascheroni constant. On using the expansion

$$\psi_0(z+1) \simeq \ln z + \frac{1}{2z} - \sum_{n=1}^{+\infty} \frac{B_{2n}}{2n\, z^{2n}} \tag{9}$$

and assuming $\Delta t \gg 1$, we can drop all terms but the logarithmic one and write:

$$\ln P(\Delta t) \simeq -\alpha\,(\gamma + \ln \Delta t) + \ln \frac{\alpha}{\Delta t}\,. \tag{10}$$

Thus, our approximated expression for the return probability is:

$$\begin{aligned} P(\Delta t) &= e^{\ln P(\Delta t)} = e^{-\gamma\alpha} \cdot \exp(-\alpha \, \ln \Delta t) \cdot \frac{\alpha}{\Delta t} = \\ &= \frac{\alpha}{\Delta t}\, e^{-\gamma\alpha} \, \exp(\,\ln\, \Delta t^{-\alpha}) = \\ &= \alpha\, e^{-\gamma\alpha}\, \Delta t^{-\alpha-1}\,, \end{aligned} \tag{11}$$

derived under the assumption that $t \gg \Delta t \gg 1$. The estimated value of $P(\Delta t)$ depends on time through α, so that the probability distribution of intervals Δt is non-stationary.

Average Return Time ($\tau = 0$)

At any given time t, the characteristic return time $< \Delta t >$ can be computed by using Eq. **4** and Eq. **11**:

$$\begin{aligned} < \Delta t > &= \sum_{\Delta t=1}^{t} \; P(\Delta t)\, \Delta t \simeq \\ &\simeq \alpha\, e^{-\gamma\alpha} \int_1^t d(\Delta t)\, \Delta t^{-\alpha} = \\ &\simeq \frac{\alpha\, e^{-\gamma\alpha}}{1-\alpha}\, t^{1-\alpha}\;. \end{aligned} \tag{12}$$

Rate Equation

Let's focus on a given word i which has frequency k_i at time t. In a continuum description, its frequency will change according to the rate equation,

$$\frac{dk_i}{dt} = (1-p)\,\Pi_i\,, \tag{13}$$

where $1-p$ is the usual probability of choosing an old word, and Π_i is the probability of picking up a previous occurrence of word i, given the times t_j of its past occurrences ($j = 1, 2, \ldots, k_i$). According to the memory kernel of Eq. **1**, the exact value of Π_i is given by the following sum over the k_i occurrences of word i:

$$\Pi_i = C \sum_{j=1}^{j=k_i} \frac{1}{\tau + (t-t_j)}\,. \tag{14}$$

Here we restrict ourselves to the case $\tau = 0$. We adopt a mean-field approach and assume that the above sum can be written as the product of the frequency k_i and the average value of the term $(t-t_j)^{-1}$ over the occurrence times t_j.

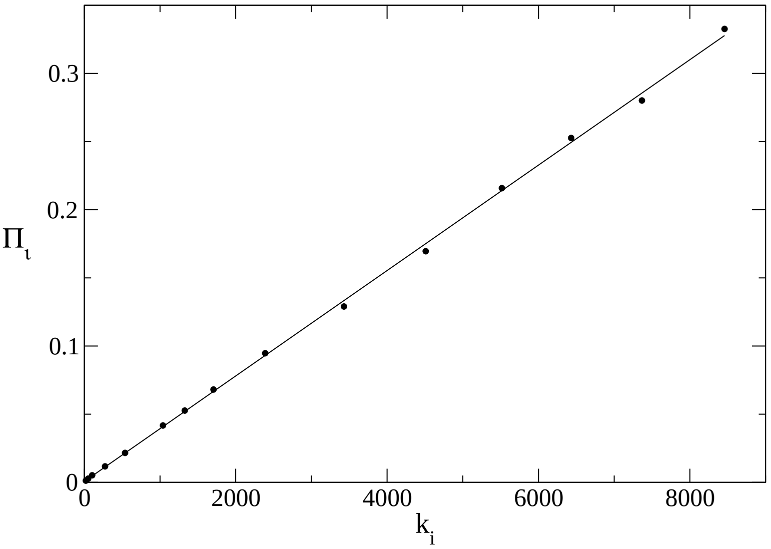

Figure 15: Rate Π_i (Eq. **14**) for a given word i having frequency k_i at time t ($p = 0.05$, $n_0 = 10$, $t = 30000$).

As shown in Fig. 15, this is supported by numerical evidence, so that we can write:

$$\Pi_i = C \sum_{j=1}^{j=k_i} \frac{1}{t-t_j} \simeq C\,k_i \left\langle \frac{1}{t-t_j} \right\rangle_j\,, \tag{15}$$

where $< >_j$ denotes the average over the k_i occurrences of word i. Furthermore, we assume that the average is dominated by the contribution of the most recent occurrence of word i, at time t_{k_i}.

$$\left\langle \frac{1}{t - t_j} \right\rangle_j \simeq \frac{1}{t - t_{k_i}}$$

We replace $t - t_{k_i}$ with the typical return interval for word i, and use Eq. **12** to estimate the latter, obtaining:

$$\left\langle \frac{1}{t - t_j} \right\rangle_j \simeq \frac{1}{t - t_{k_i}} \simeq \frac{1}{< \Delta t >} = \frac{1-\alpha}{\alpha\, e^{-\gamma\alpha}} \cdot \frac{1}{t^{1-\alpha}}\,, \tag{16}$$

which has a (sublinear, as $\alpha > 0$) power-law dependence on t and a slower time dependence through α. Fig. 16 shows that the above expression captures the correct temporal dependence of the average $< t - t_j >^{-1}$ for a given frequency k_i, provided that a constant factor Ω is introduced, as follows:

$$\left\langle \frac{1}{t - t_j} \right\rangle_j \simeq \frac{1}{\Omega} \cdot \frac{1-\alpha}{\alpha\, e^{-\gamma\alpha}} \cdot \frac{1}{t^{1-\alpha}}\,. \tag{17}$$

The need for a corrective factor Ω is a consequence of our simplifying assumptions, namely our mean-field approximation, the fact that we ignored all occurrences of word i but the very last, and the approximations underlying our estimate of the return time Δt. Moreover, as shown in Fig. 16, Ω shows a weak dependence on the frequency k_i of the selected word i, especially for small values of k_i. In order to keep only the linear dependence of the kernel on k_i we approximate Ω with its average value over k, numerically estimated as $\Omega \simeq 1.61$ (see Fig. 1). While this is certainly a rather crude approximation, it appears to work remarkably well, as we will show later.

We introduce Eq. **17** and Eq. **15**, into the rate Eq. **13**, obtaining:

$$\frac{dk_i}{dt} \simeq (1-p) C\, k_i \left\langle \frac{1}{t - t_j} \right\rangle_j = \frac{k_i}{\Omega} \cdot \frac{1-\alpha}{e^{-\gamma\alpha}} \cdot t^{\alpha-1}\,. \tag{18}$$

We integrate Eq. **18**, with the assumption of considering α constant, from time t_i, when word i appeared for the first time (with frequency 1) the the final time t, when word i has frequency k_i,

$$\int_1^{k_i} \frac{dk_i'}{k_i'} = \frac{1-\alpha}{\Omega\, e^{-\gamma\alpha}} \cdot \int_{t_i}^{t} dt'\, t'^{\alpha-1}\,. \tag{19}$$

Performing the integration we get

$$\ln k_i = \frac{1-\alpha}{\Omega\, \alpha\, e^{-\gamma\alpha}} \left(t^\alpha - t_i^\alpha\right)\,, \tag{20}$$

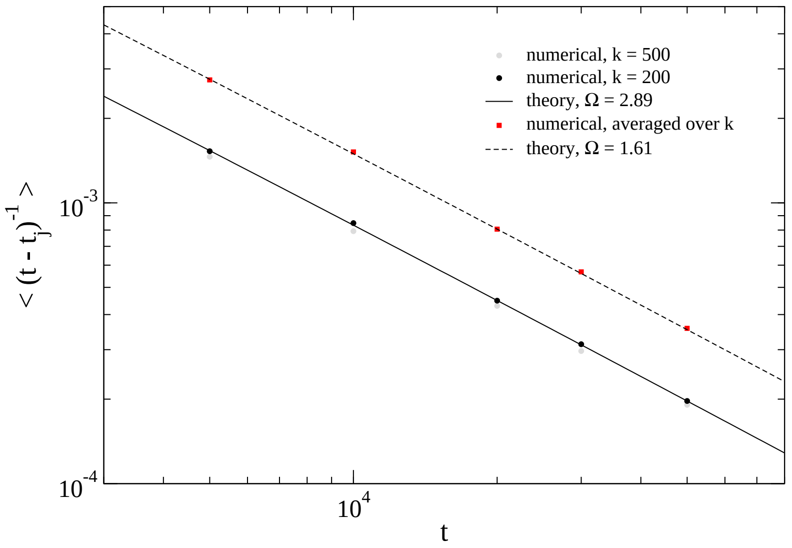

Figure 16: Memory kernel of Eq. **17** averaged over the times of occurrence t_j and over about 2000 realizations of the process ($p = 0.05$, $n_0 = 10$, $t = 5 \times 10^3, 10^4, 2 \times 10^4, 3 \times 10^4, 5 \times 10^4$). Values are shown for a word of given frequency $k = 200$ (black dots), a word of frequency $k = 500$ (gray dots) and the averaged over all frequencies (red dots, above). Numerical error bars are within the size of data markers.

which can be written as

$$k_i = \exp\left[\frac{1-\alpha}{\Omega\,\alpha\,e^{-\gamma\alpha}}\,t^{\alpha}\right] \cdot \exp\left[-\frac{1-\alpha}{\Omega\,\alpha\,e^{-\gamma\alpha}}\,t_i^{\alpha}\right] = A\,e^{-Kt_i^{\alpha}}\ , \tag{21}$$

where we defined

$$K \equiv \frac{1-\alpha}{\Omega\,\alpha\,e^{-\gamma\alpha}}\ ,\ A \equiv e^{Kt^{\alpha}}\ . \tag{22}$$

Eq. **21** shows that k_i has a stretched exponential dependence on t_i. On solving it for t_i, we can define a characteristic time of appearance t^* for a word that has frequency k at time t,

$$t^*(k,t) = \left[\frac{\ln(A/k)}{K}\right]^{1/\alpha}\ . \tag{23}$$

We have dropped the index i, since both k and t refer no longer to a specific word i (and Ω no longer refers to any specific word, as already mentioned).

Probability Distribution of Frequencies

At time t, the fraction of word frequencies less than k is given by the cumulated distribution $P_<(k)$. Let us define $P_>(t^*(k,t))$ as the probability of observing an appearance time in excess of $t^*(k,t)$, and $P_<(t^*(k,t)) = 1 - P_>(t^*(k,t))$. We have

$$P_<(k) = P_>(t^*(k,t)) = 1 - P_<(t^*(k,t))\,. \tag{24}$$

The probability $P_<(t^*(k,t))$ is equal to the fraction of words which appeared earlier than $t^*(k,t)$: since we know that a new word appears with probability p per unit time, the number of words that appeared earlier than $t^*(k,t)$ is simply $pt^*(k,t)$, and their relative fraction is

$$P_<(t^*(k,t)) = \frac{p\,t^*(k,t)}{n_0 + pt}\,. \tag{25}$$

The probability distribution for word frequencies $P(k)$ can be computed as

$$P(k) = \frac{\partial P_<(k)}{\partial k} = \frac{p}{(n_0 + pt)\,(K\alpha)\,k}\left[\frac{\ln(A/k)}{K}\right]^{\frac{1}{\alpha}-1}\,, \tag{26}$$

and is in good agreement with numerical evidence, as shown in Fig. 17.

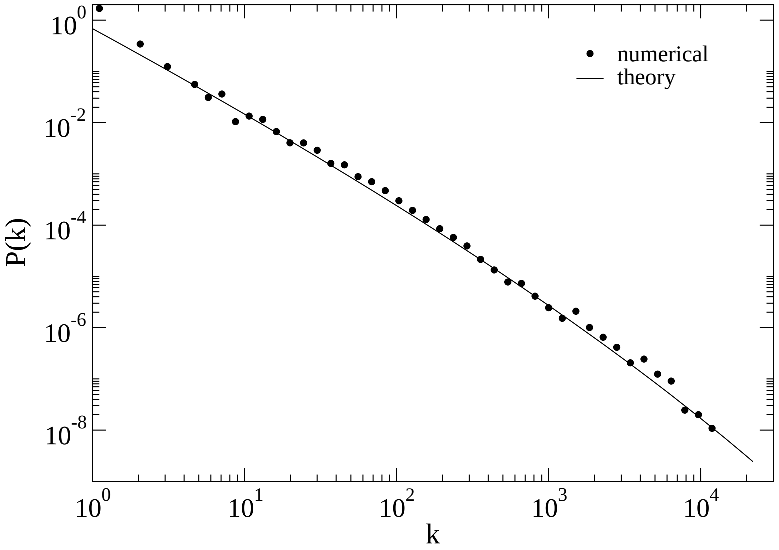

Figure 17: Probability distribution $P(k)$ for the frequency of word occurrence. Numerical data (dots, average over 50 realizations) are in excellent agreement with Eq. **26** (solid line) ($p = 0.05$, $n_0 = 10$, $t = 30000$, $\Omega = 1.61$).

Ranked Distribution of Frequencies

The rank $R(k')$ of a word with frequency k' can be written in terms of the frequency probability distribution as

$$R(k') \simeq (n_0 + pt) \int_{k'}^{k_{\max}} P(k)\, dk \ .$$

Using Eq. **26** for $P(k)$ and Eq. **24** for $P_<(k')$, we write:

$$\begin{aligned} R(k') &= (n_0 + pt) \int_{k'}^{k_{\max}} \frac{\partial P_<(k)}{\partial k}\, dk = \\ &= (n_0 + pt)\ [1 - P_<(k')] = p\, t^*(k', t)\,, \end{aligned} \tag{27}$$

showing that in our treatment of the process, the ratio R/p plays the role of a characteristic time of arrival for a word. Inserting Eq. **23** into the above equation, we get

$$R(k) \simeq p \left[\frac{\ln(A/k)}{K} \right]^{1/\alpha} \,, \tag{28}$$

and the ranked frequency distribution is a stretched exponential in R/p,

$$k(R) \simeq A \exp \left[-K \left(\frac{R}{p} \right)^{\alpha} \right] . \tag{29}$$

This can be normalized dividing by $n_0 + t$, the total number of words at time t, finally yielding the probability density for word rank R:

$$P(R) \simeq \frac{A}{n_0 + t} \exp \left[-K \left(\frac{R}{p} \right)^{\alpha} \right] . \tag{30}$$

Fig. 18 shows that the above equation is in fair agreement with numerical evidence. Moreover, all the ranked distributions we observed for $\tau = 0$ can be reproduced accurately by a stretched exponential of the above form. It should be noticed that the agreement in Fig. 18 is obtained without any explicit fitting procedures since the value of the only free parameter Ω has been chosen with the averaging procedure described above.

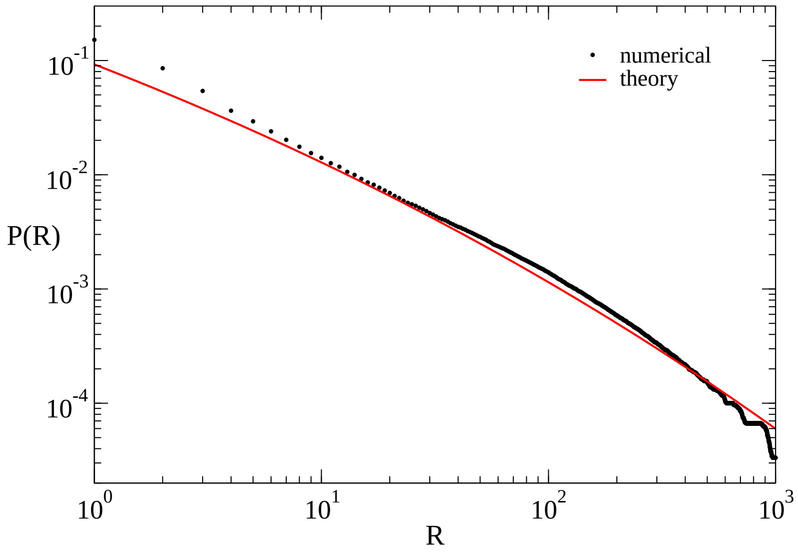

Figure 18: Ranked probability distribution $P(R)$. Numerical data (dots, average over 50 realizations) are compared against the prediction of Eq. **30** (solid line) ($p = 0.05$, $n_0 = 10$, $t = 30000$, $\Omega = 1.61$).

Conclusions

Before concluding several remarks are in order.

- We have presented a preliminary continuum approach to explain the behaviour of the Yule-Simon microscopic model with memory proposed in the paper. The presence of a long-term memory kernel makes the rigorous treatment non trivial. Our approach makes a certain number of assumptions, sometimes rough (but checked numerically), especially to guess the functional form of the memory kernel both as a function of time and of the frequency of a word. Nonetheless this approach allows for an excellent agreement between theory and simulation for the frequency probability distribution $P(k)$. This is somehow the signature that it is capturing some essential features of the original microscopic model.

  Our approach requires a single phenomenological parameter (Ω), for which we have no theoretical estimates, at present. The rank probability distribution $P(R)$ appears to be much more sensitive to the approximations we made, but the agreement between numerics and theory is nevertheless reasonable.

  For $\tau = 0$ no power-laws are observed in the tails of neither $P(k)$ or $P(R)$, but rather we observe (both numerically and analytically) a slowly varying slope in log-log plots.

- Of course this analysis is still preliminary and it represents only a first step towards a deeper comprehension of our model. In particular it will be important to have a stronger control of the approximations made especially for what concerns all the quantities (like α) slowly varying with time.

  The whole approach can be in principle extended to the $\tau \neq 0$ and work is presently in progress to understand the role of this parameter. A lot of questions arise, for instance: (I) what is the role of τ on long-time scales; (II) does it only affect the dynamics on short time-scales mimicking the effect of a Yule-Simon model without memory? (III) does the limit $\tau \simeq t$ falls in the same universality class of the Yule-Simon model without memory?